AI Ethics

Francesca Rossi

IBM AI Ethics Global Leader

Politecnico di Milano, Sept. 19th, 2019

What is AI?

Intelligence and rationality

A scientific discipline that aims to create machines (hw/sw) that show a behavior that would be called intelligent if seen in a human being

Rationality

Given a problem, to know (and act) how to best solve it

Narrow and general AI

Narrow AI

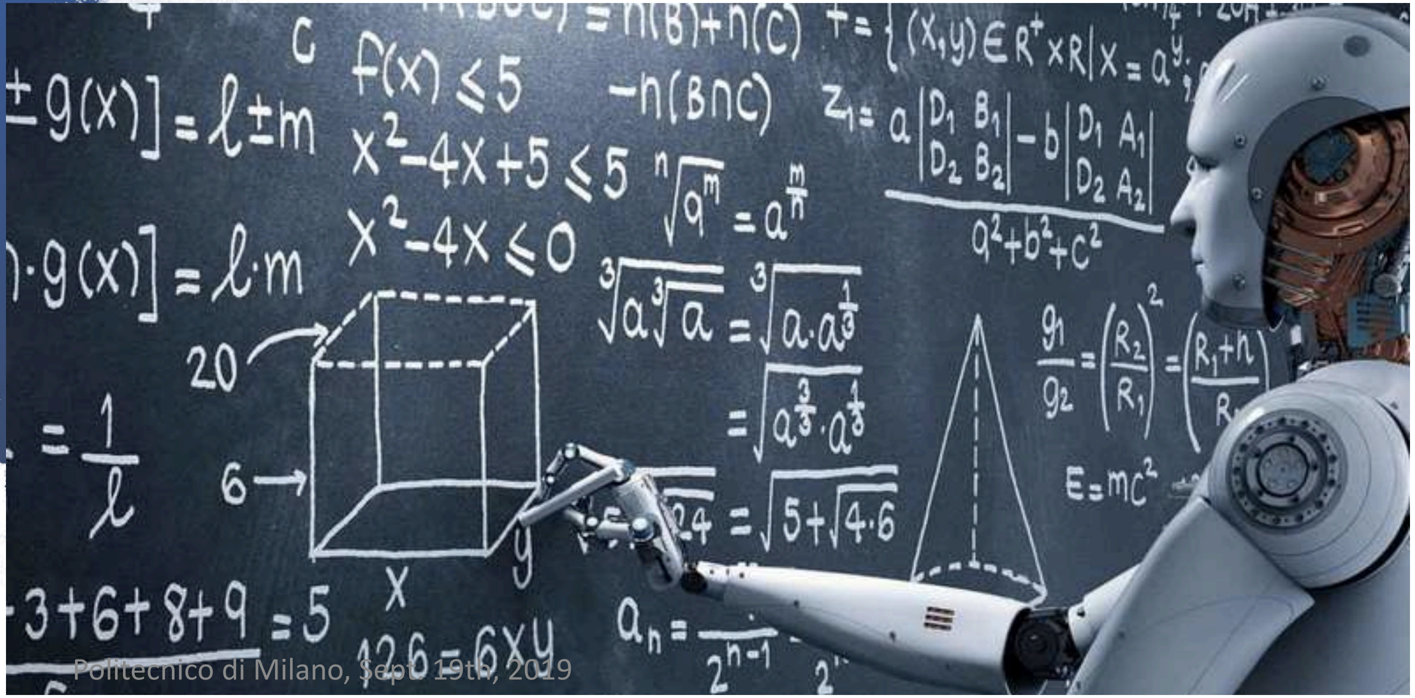
Can solve specific problems

Vertical and specific

General AI

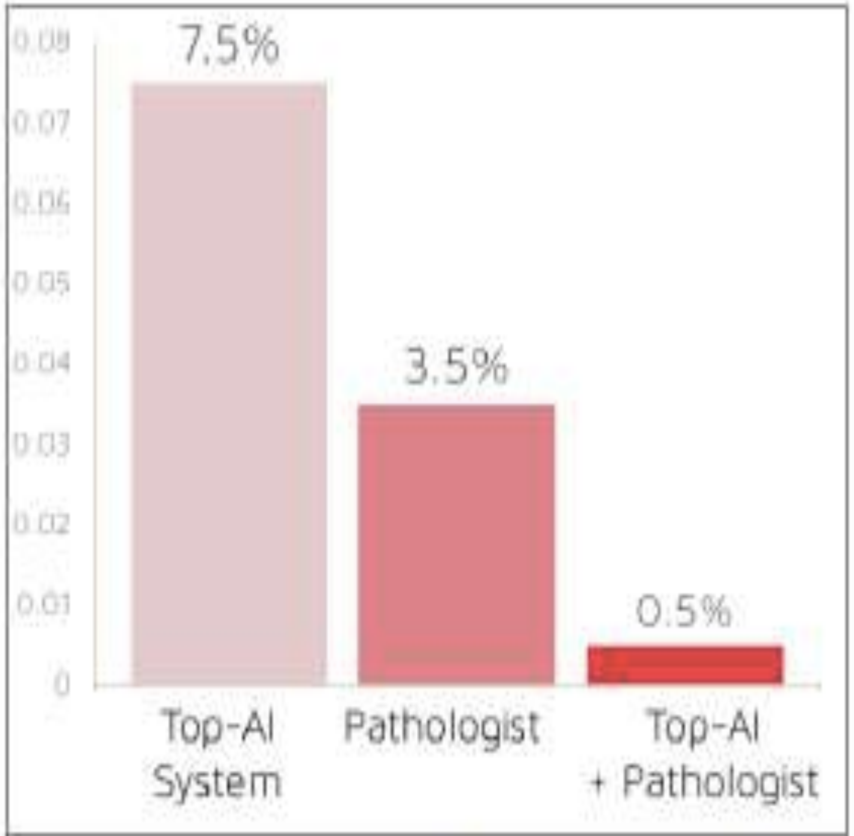
Can handle different problems and scenarios

Horizontal



Artificial Intelligence vs Augmented Intelligence

Human-machine collaboration
Complementarity



AI significantly reduces pathologist error rate in the identification of metastatic breast cancer from sentinel lymph node biopsies.

How does it work?

How to teach a machine how to solve a problem?

1. Logical reasoning

   We tell it the steps to be done to solve the problem

2. Machine learning

   We give examples of problem's solutions and we provide methods to generalize

AI, Machine Learning, Deep Learning

ARTIFICIAL INTELLIGENCE

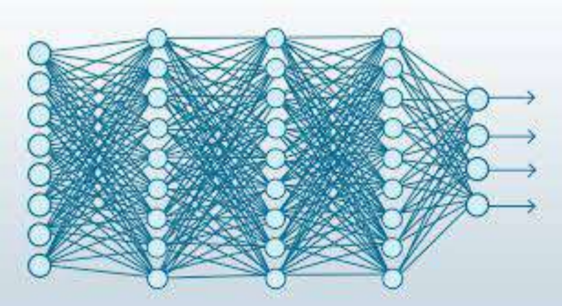
Intelligent algorithms defined and coded by people into machines

MACHINE LEARNING

Ability to learn without being explicitly programmed

DEEP LEARNING

Learning based on Deep Neural Networks



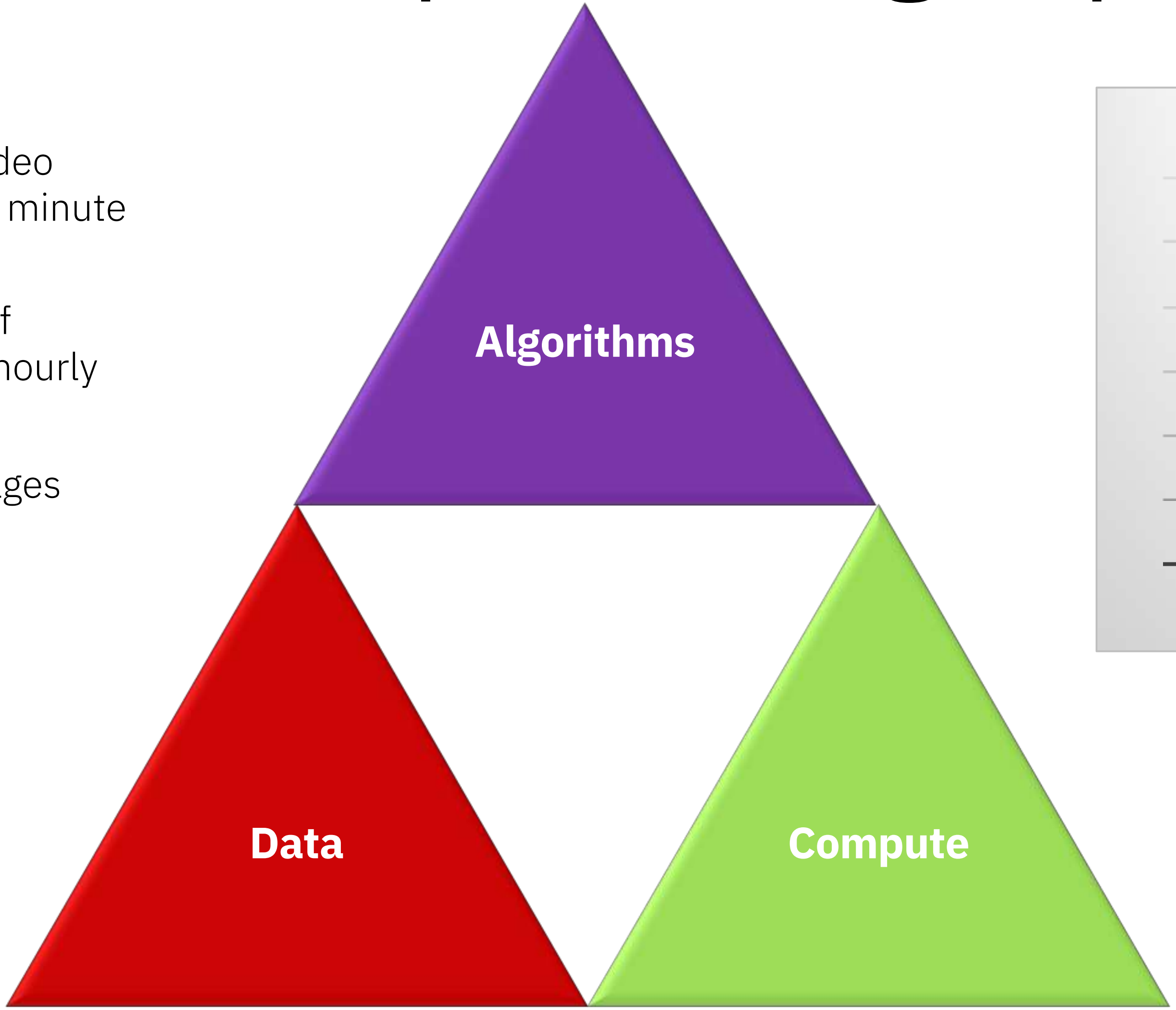
1950's | 1960's | 1970's | 1980's | 1990's | 2000's | 2006's | 2010's | 2012's | 2017's

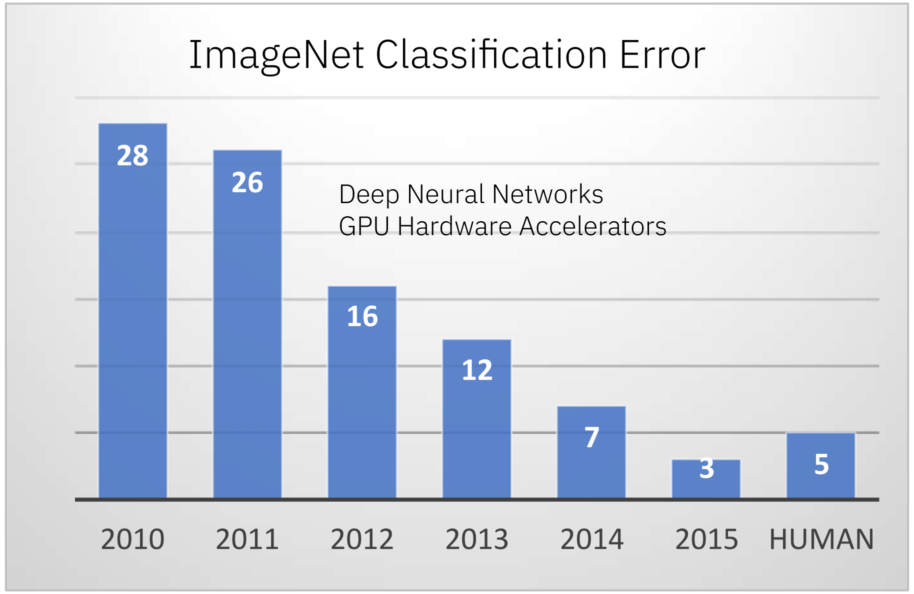
Machine/Deep Learning explosion

YouTube
400 hours of video uploaded every minute

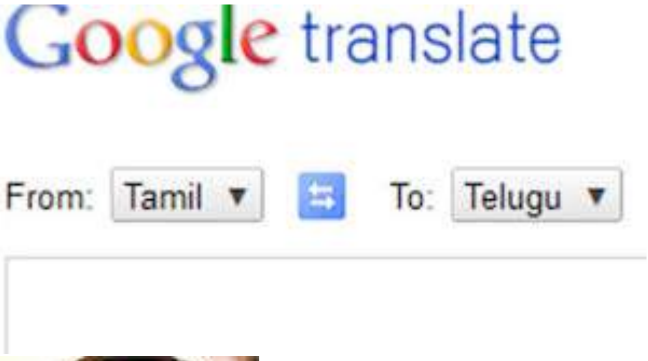
Walmart
2.5 petabytes of customer data hourly

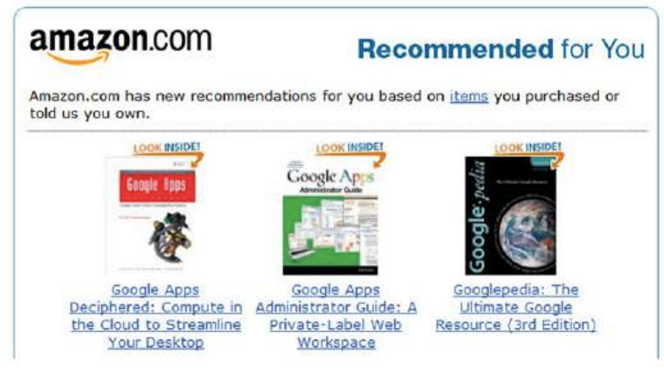
Facebook
350 million images uploaded daily

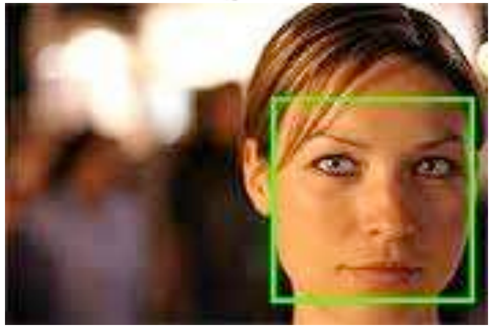


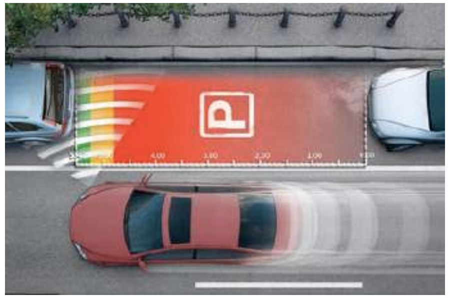


AI in our life





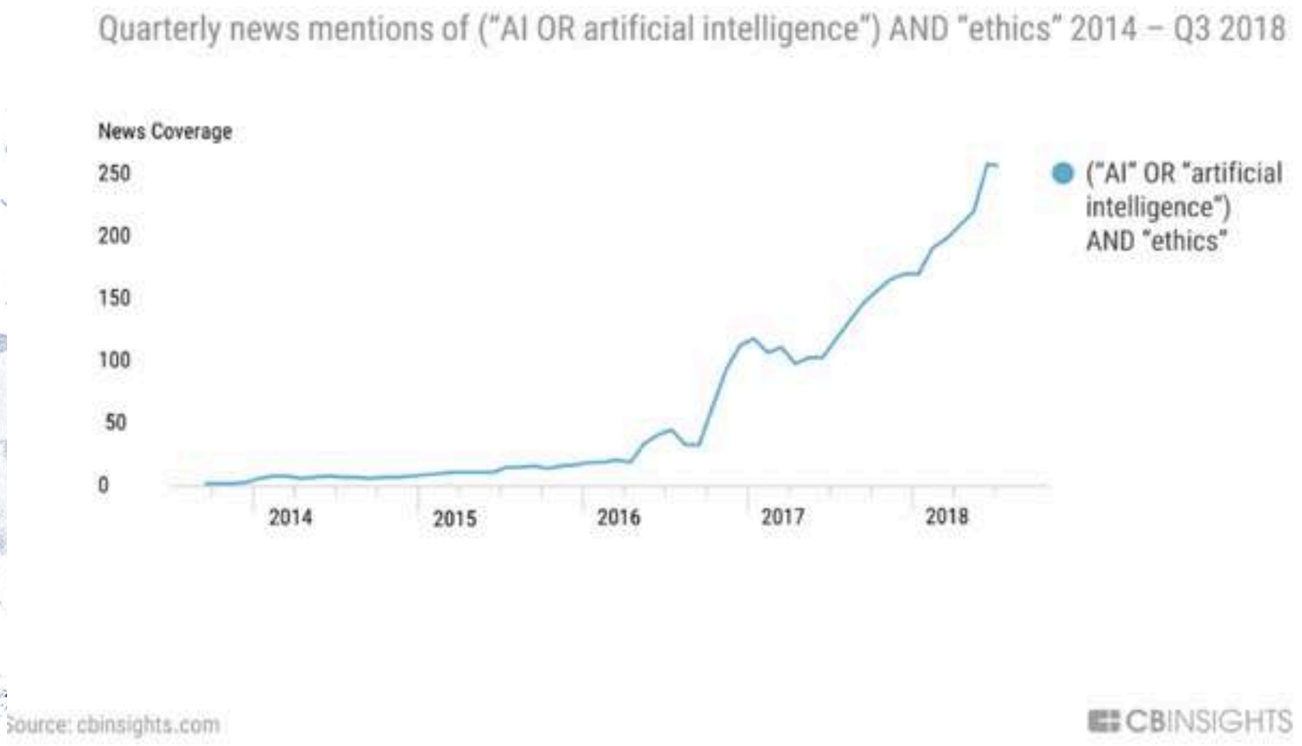




Current limitations

- Common sense reasoning
- Combination of learning and reasoning
- Natural language understanding
- Learning from few examples
- Learning general concepts
- Robustness/adversarial examples

AI ethics



Quarterly news mentions of ("AI OR artificial intelligence") AND "ethics" 2014 – Q3 2018
News Coverage
250
200
150
100
50
0
2014
2015
2016
2017
2018
("AI" OR "artificial intelligence") AND "ethics"
Source: cbinsights.com
CBINSIGHTS

Main concerns

Bias

Careful with the examples!

Value alignment

Ethical, moral, social, and legal constraints

Black box

Must be able to explain its decisions

properties of the technology

Data issues

Privacy, storage, ownership, use

Accountability

Who is responsible if something goes wrong?

governance and rules

Impact on jobs

How to cope with job transformation?

Impact on society

People-machine and people-people interactions

Deep fake

AI can generate content that looks real but it is not

societal issues

Autonomous weapons, surveillance systems

Are these acceptable uses?

possible vs desirable uses

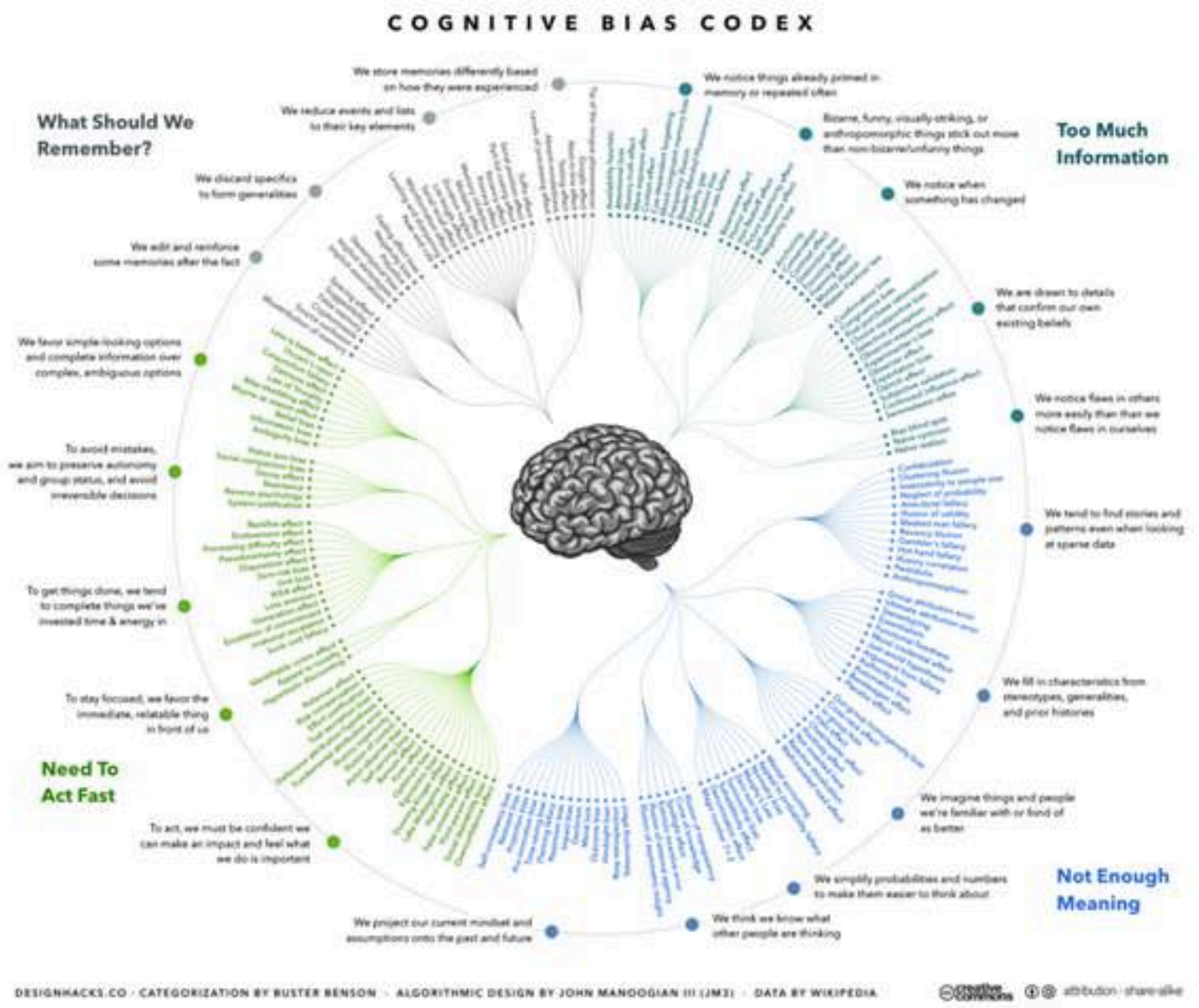
Superintelligence

Will we lose control?

long term concerns

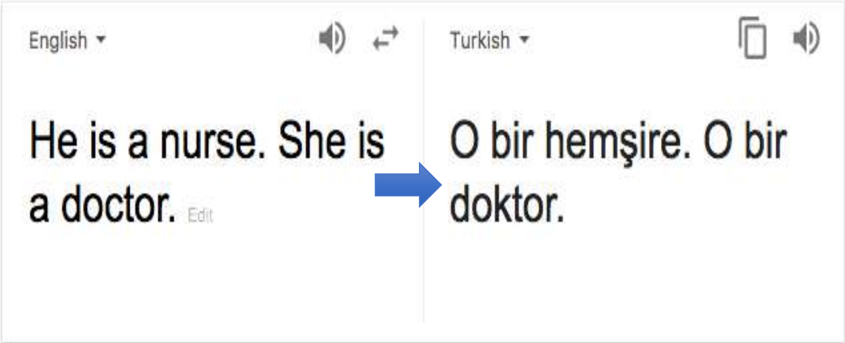
AI bias

Humans are bias

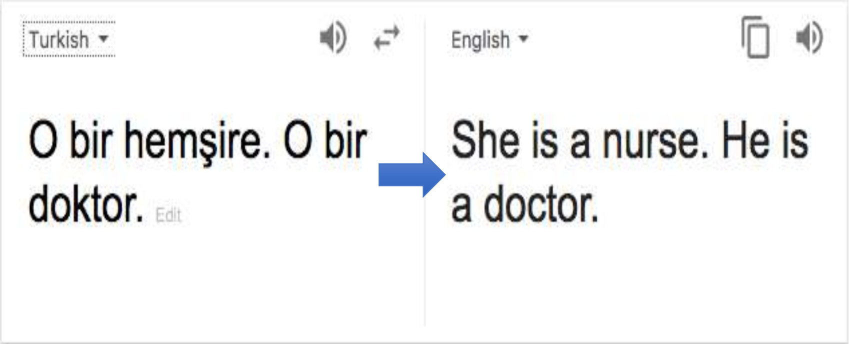


AI bias

From biased training data to unfair decisions



English to Turkish



Turkish to English

What is being done?

- **Technological solutions**
  - Bias, explainability, security
- **Other tools**
  - Principles, guidelines, best practices, incentives, standards, certificates, audits, laws, …
- **Enablers**
  - Education and dissemination
    - AI students, developers, users, impacted communities, policy makers
  - Governance
    - Within each AI company and globally
- **Multi-disciplinary efforts**
  - AI, sociology, psychology, economy, philosophy, law

What is being done?

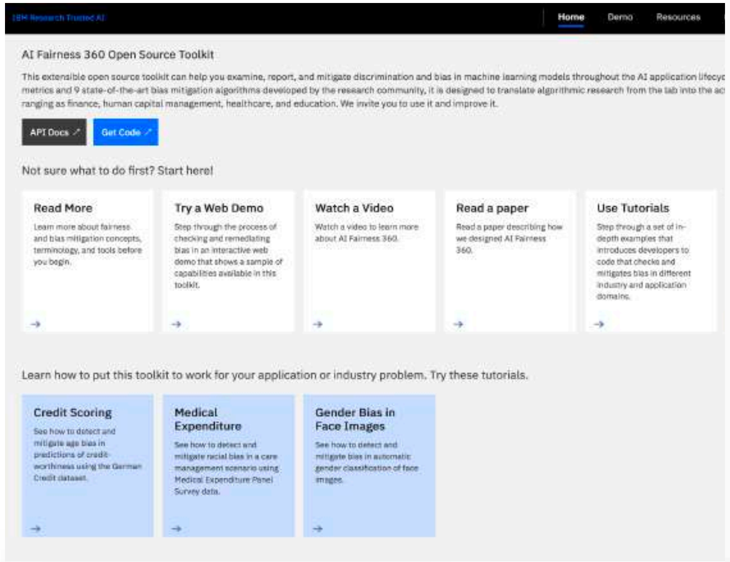
Education

COURSE TITLE	CODE	UNIVERSITY	DEPARTMENT
Algorithms and Society	EECS 395 and COMMST 395	Northwestern University	Computer Science and Communication Studies
Code and Power	LIS 500	University of Wisconsin - Madison	Information School
Designing Field Experiments at Scale	SOC 412	Princeton University	Sociology, Center for IT Policy
Digital Anthropology	ANTH 4020	University of Colorado Boulder	Continuing Education
Ethical and Policy Dimensions of Information, Technology, and New Media	INFO 4601/5601	University of Colorado Boulder	Information Science
Ethical and Social Implications of Data		Marquette University	Computer Science
Ethics in Business Analytics	ITAO 40510	University of Notre Dame	IT, Analytics, & Operations
Ethics in Data Science		University of Utah	Computer Science

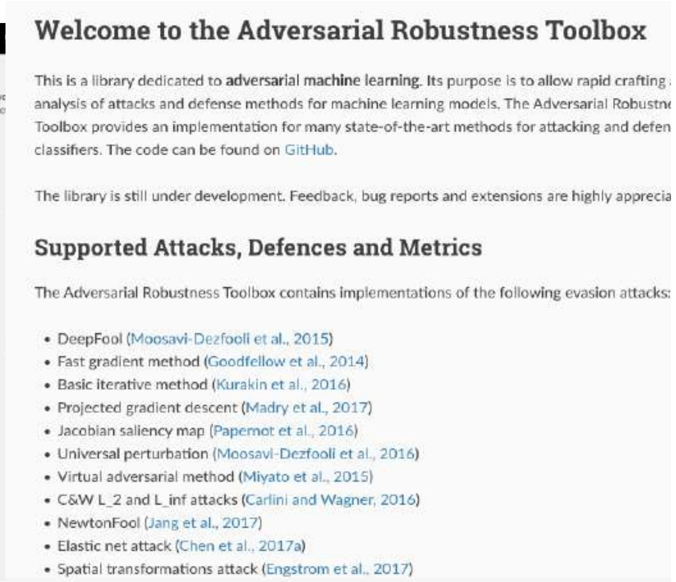
Crowdsourced list of AI/CS ethics courses (238 entries so far), Casey Fiesler, CU Boulder

AI Ethics at IBM

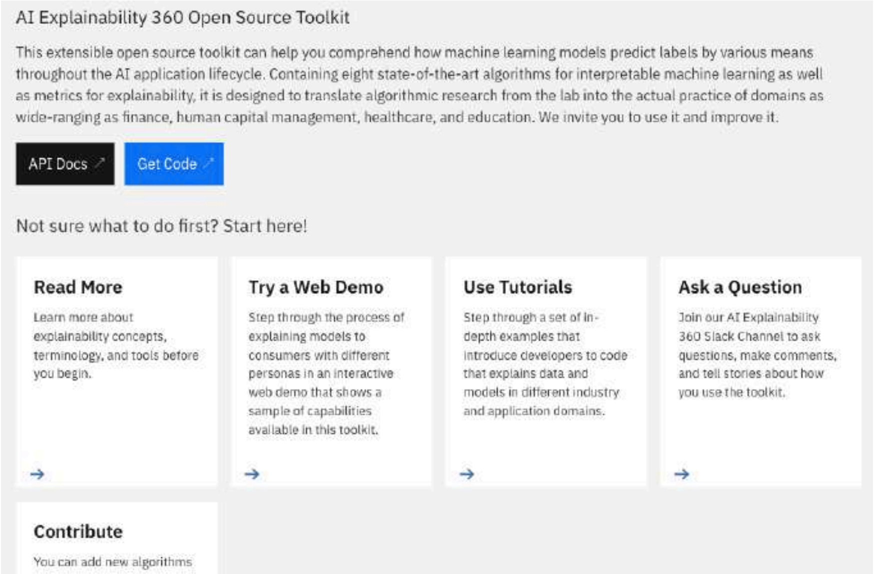
AI Fairness 360



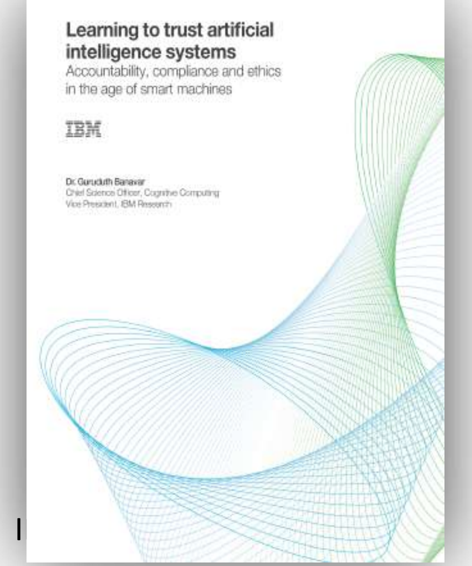
Adversarial Robustness Toolbox



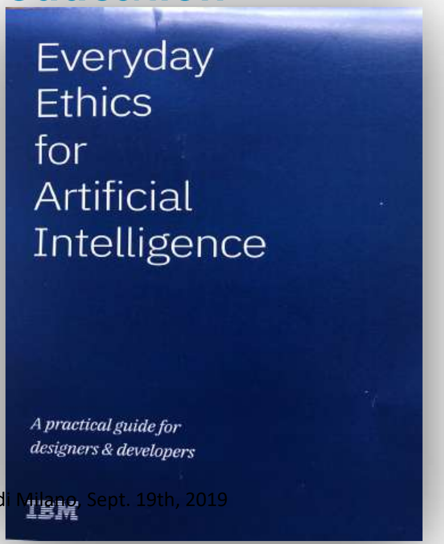
AI Explainability 360



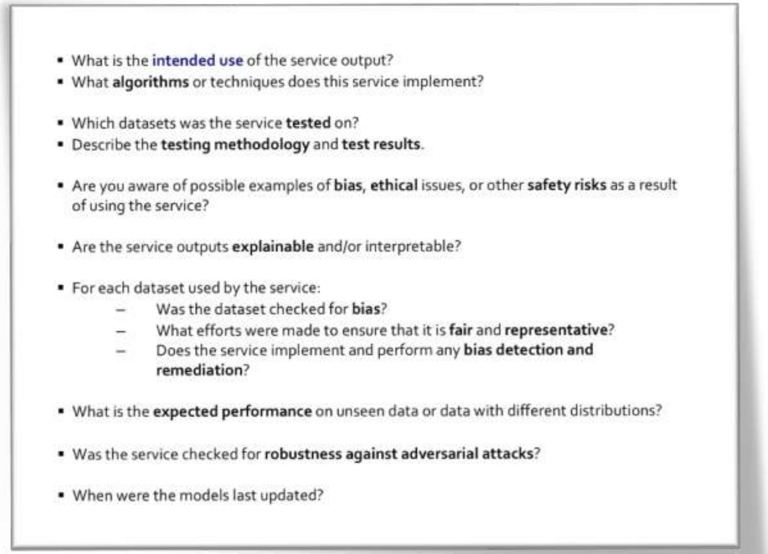
2016 IBM white paper



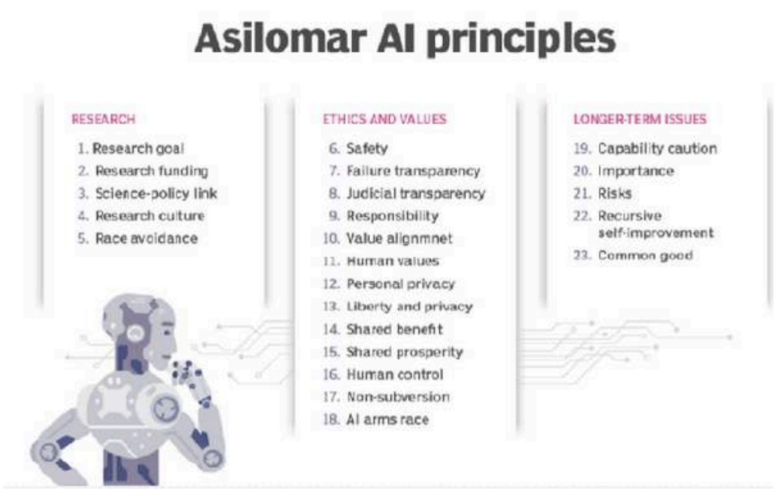
AI ethics developer education



AI factsheets

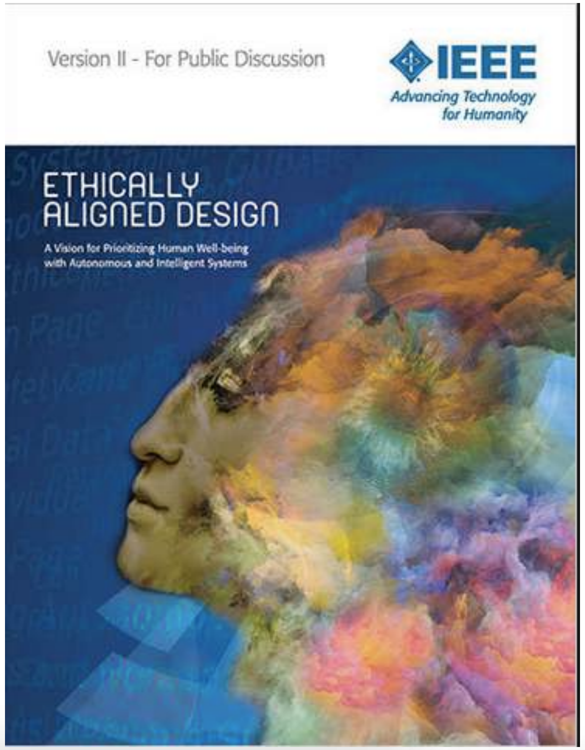


- What is the **intended use** of the service output?
- What **algorithms** or techniques does this service implement?
- Which datasets was the service **tested** on?
- Describe the **testing methodology** and **test results**.
- Are you aware of possible examples of **bias**, **ethical** issues, or other **safety risks** as a result of using the service?
- Are the service outputs **explainable** and/or interpretable?
- For each dataset used by the service:
  - Was the dataset checked for **bias**?
  - What efforts were made to ensure that it is **fair** and **representative**?
  - Does the service implement and perform any **bias detection and remediation**?
- What is the **expected performance** on unseen data or data with different distributions?
- Was the service checked for **robustness against adversarial attacks**?
- When were the models last updated?



WORLD ECONOMIC FORUM

AI4PEOPLE

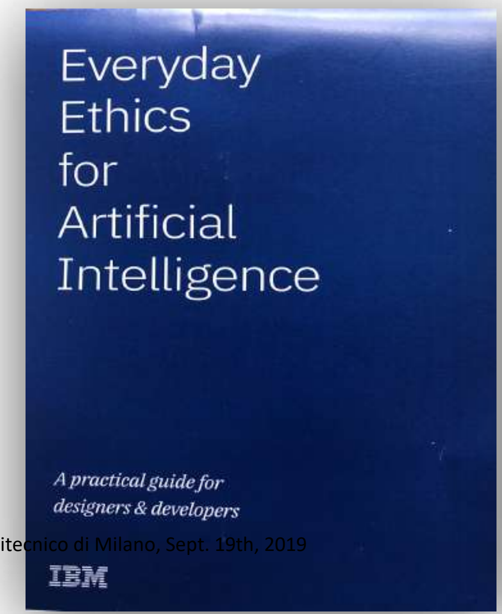


AI for Good Global Summit
An ITU experience

future of life INSTITUTE

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY



Embedding ethics into decision support systems

Personalization vs (ethical) behavioral constraints

Preferences are essential to allow for personalized services

Online recommendations, healthcare, financial advisors, etc

Also a way to tell a system what we want from it

But they need to be combined with other priorities to avoid undesired actions

Ethics principles, moral values, business guidelines, behavioral constraints, common sense reasoning

Value alignment

To achieve a goal, given by a human
in the best way
possibly being creative and innovative
while being aligned to the appropriate values for the task

Ethically bounded AI

- How do we bound the behavior of autonomous agents, without explicitly telling them what to do, in a way that it will achieve the goal while complying with appropriate ethical/behavioral constraints?

- Two main approaches:
  - Top Down: write down all the rules and have the agent follow them
    - We need to know the best strategy to solve the problem
  - Bottom Up: show the agent appropriate actions
    - Data-driven approach

Two explored solutions

1. Recommendation systems

- Goal: to teach AI systems how to obey behavioral constraints learned by observation while still being responsive to the feedback from users
  - Reinforcement Learning approach
  - Examples to describe the ethical constraints, learnt offline
  - Constrained RL behavior during online use

2. Preferences and ethical priorities

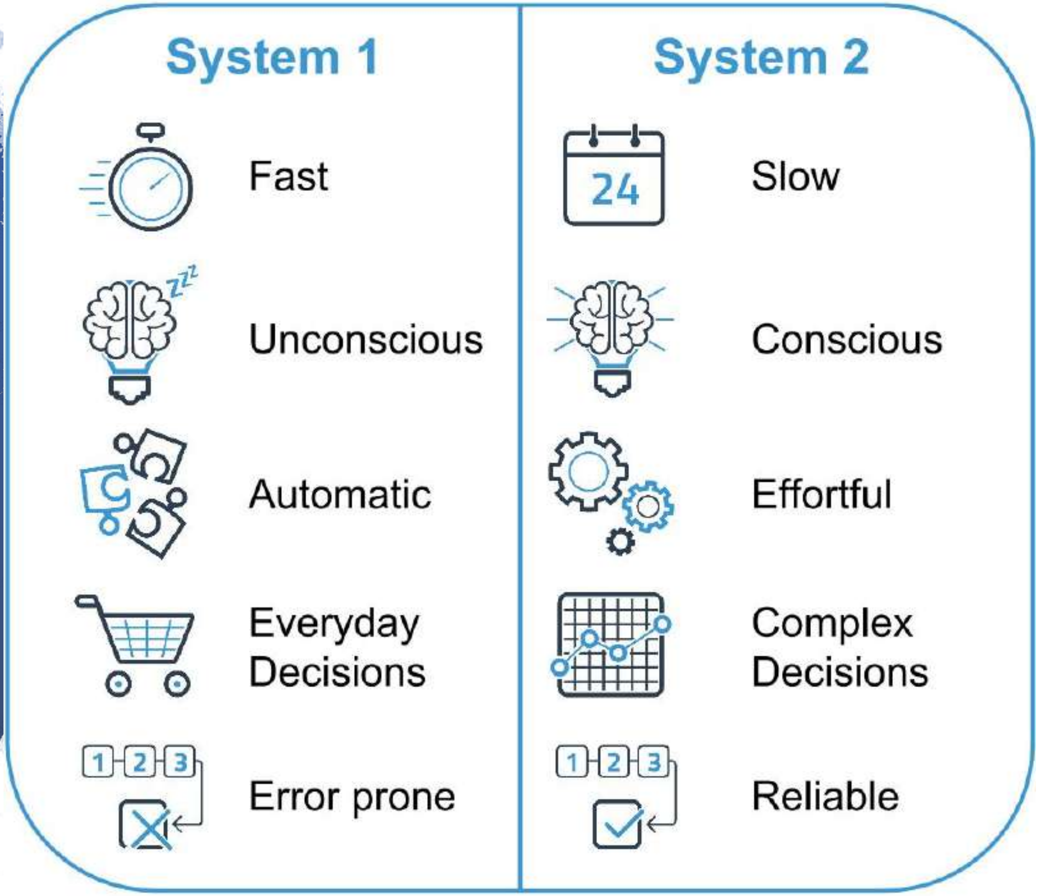
- Goal: To achieve personalization while not compromising essential values and principles
  - Preference frameworks (CP-nets) to model both preferences and ethical guidelines
  - Distance between CP-net structures
  - Distance thresholds to decide if agent can follow its preferences or must be better aligned to ethical priorities

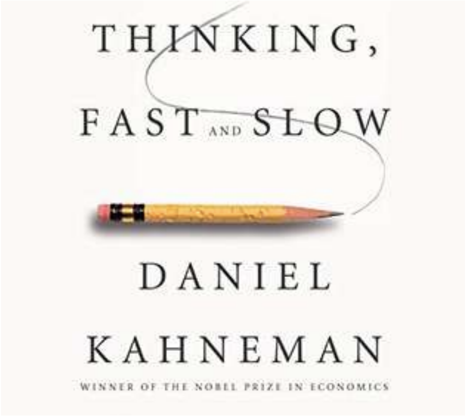
Conference papers: AAAI 2018, AAMAS 2018, AIES 2018
Book chapters:

- *Artificial Intelligence Safety and Security, CRC Press, 2018*
- *Ethics of AI, Oxford University Press, 2019*

Which is better?

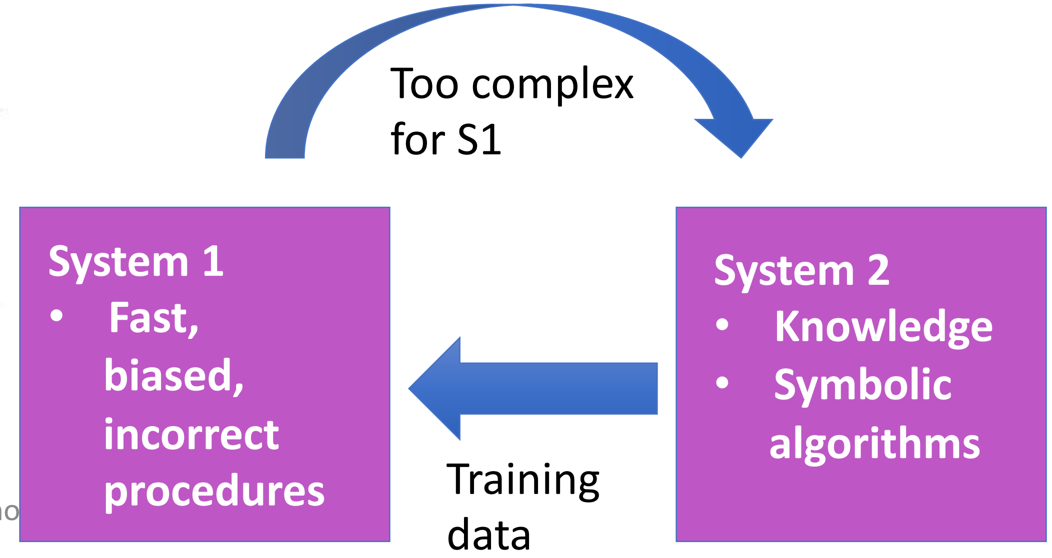
Should we choose?
What happens in our brain?





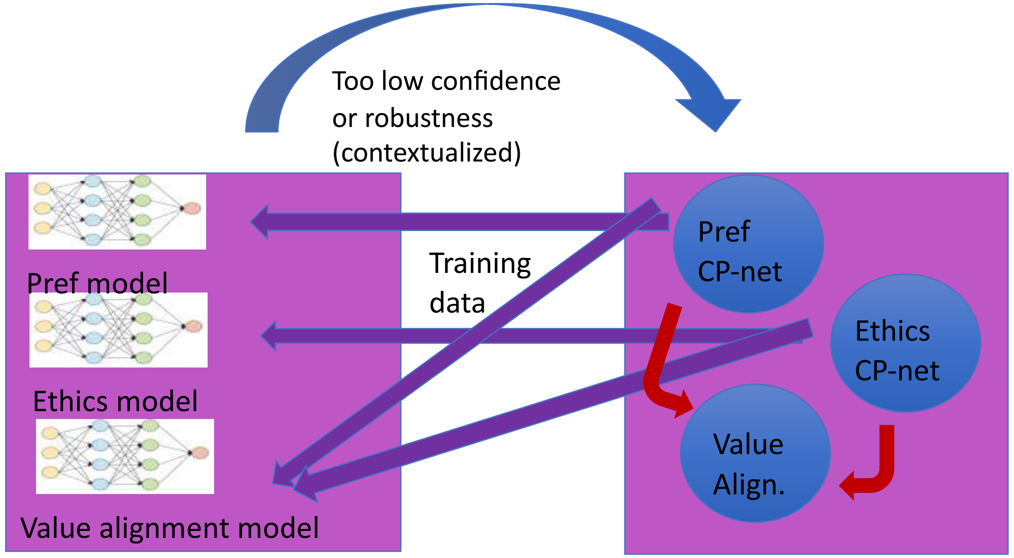
Fast and slow AI

- System 1 resembles a data-driven approach
  - Training data provided over time by System 2
  - After a while, some tasks pass to system 1, while others always require system 2
  - Ex.: reading a word, arithmetic 2-digit multiplication
- System 2 resembles a symbolic/logic approach
  - Understands how to tackle new or computationally difficult tasks
  - Computational complexity is (one of) the trigger(s) for system 2



Fast and slow preferences and ethical constraints

- A system 1 and a system 2 version for both preferences and ethical constraints
- The Pref/Ethics/VA modes answer to
  - What is the most preferred/ethical choice?
  - Is choice A dominated by choice B?
  - How far are my preferences from the ethical constraints?
- But only after he got enough training data from the Pref/Ethics/VA symbolic/logic procedure



Conferences: AAMAS 2019

What next?

- Implement the full dual-agent architecture
- Understand when S1 should call S2
  - or when S2 should awake and override S1
- Contextualize to tasks and scenarios

- Where to place causality and common sense reasoning?

- More than one ethical theory
  - Deontology: constraints and priorities
  - Utilitarianism
  - Contractualism

- Elephant in the Room: Where do values and ethical constraints come from?
  - Multi-stakeholder approach

Thanks!